Processus de tests de rapports de vraisemblance pour la détection de QTL

Charles-Elie Rabier, Jean-Marc Azaïs, Céline Delmas

Processus de tests de rapport de vraisemblance pour la détection de QTL

Charles-Elie Rabier[1,2] & Jean-Marc Azaïs[1] & Céline Delmas[2]

[1] *Université de Toulouse,*
Institut de Mathématiques de Toulouse, U.P.S. et I.N.S.A,
F-31062 Toulouse Cedex 9
charles-elie.rabier@insa-toulouse.fr, azais@cict.fr
[2] *Station d'Amélioration Génétique des Animaux*
INRA, Auzeville B.P. 52627, 31326 Castanet Tolosan
celine.delmas@toulouse.inra.fr

Résumé : On considère le processus de tests de rapport de vraisemblance (LRT) en référence au test d'absence de QTL sur un intervalle $[0, T]$ représentant un chromosome (QTL désigne un gène à effet quantitatif). On étudie la distribution asymptotique du processus de LRT sous l'hypothèse nulle d'absence de QTL sur $[0, T]$, et sous l'alternative générale qu'il existe m QTL sur $[0, T]$. On suggère d'estimer le nombre de QTL, leurs positions, et leurs effets par vraisemblance pénalisée. Les résultats seront généralisés au cas où les individus sont structurés en familles.

Abstract : We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL on the interval $[0, T]$ representing a chromosome (a QTL denotes a quantitative trait locus, i.e. a gene with quantitative effect on a trait). We give the asymptotic distribution of this LRT process under the null hypothesis that there is no QTL on $[0, T]$ and under the general alternative that there exist m QTL on $[0, T]$. We propose to estimate the number of QTL, their positions and their effects by penalized likelihood. Our results are extended to the case where individuals are structured into families.

Mots-clés : Génome, Détection de QTL, Processus Gaussiens, Modèles de mélange.

Keywords : Genome, QTL Detection, Gaussian processes, Mixture models.

1 Motivation

Les nouvelles technologies en matière de génomique se révèlent être efficaces afin de percer les secrets de la variation génétique d'un caractère quantitatif. Ces technologies permettent la caractérisation moléculaire de marqueurs polymorphes (i.e. présentant plusieurs allèles) sur l'ensemble du génome. Ces derniers seront par la suite utilisés pour identifier et localiser les loci (i.e. emplacements physiques précis sur un chromosome) où la variation allélique est associée à la variation du caractère quantitatif considéré. On

nomme QTL de tels loci.

Cependant, la détection et la localisation de QTL s'avère difficile d'un point de vue statistique. On se propose ici d'étudier la technique statistique qui consiste à scanner le génome, nommée "Interval Mapping", proposée par Lander et Botstein (1989).

2 L'étude dans sa globalité

L'étude porte sur une population de descendants d'un père. La problèmatique se veut la détection d'un QTL sur un chromosome. Le phénotype est observé sur n individus. On note Y_j, $j = 1, .., n$, les observations que l'on supposera indépendamment et identiquement distribuées (iid). Le mécanisme de la génétique implique que parmi les deux chromosomes de chaque individu, l'un est herité de la mère (son effet sera négligé) et l'autre du père. Celui transmis par le père est constitué, en raison de crossing-over, de parties provenant du chromosome 1 du père et de parties provenant du chromosome 2 du père.
Une population back-cross, $A \times (A \times B)$, où A et B sont de pures lignées homozygotes, est un cas particulier de la population étudiée. A l'aide de la distance et de la modélisation de Haldane (1919), chaque chromosome sera représenté par un segment $[0, T]$. La distance sur $[0, T]$ est appelée distance génétique (mesurée en Morgans). Le point clé est que, si la vraie position du QTL est $t = t^\star$, la réponse Y obéit à un modèle de mélange dont les poids sont connus :

$$p_t f_{(\mu+q,\sigma)}(.) \; + \; (1 - p_t) f_{(\mu-q,\sigma)}(.) \tag{1}$$

où $f_{(\mu,\sigma)}(.)$ désigne une densité Gaussienne de moyenne μ et de variance σ^2. (μ, q, σ) sont les paramètres inconnus. A chaque position $t \in [0, T]$, est effectué un test du rapport de vraisemblance (LRT) de l'hypothèse "q=0" dans la formule (1) basé sur n observations Y_1, ..., Y_n. La quantité obtenue est notée Λ_t. Le processus $\Lambda_{(.)}$ est appelé processus de tests de rapport de vraisemblance. Le choix comme statistique de test du maximum de ce processus, revient à effectuer un LRT dans un modèle où la localisation du QTL est un paramètre supplémentaire.

Dans le cas particulier où les poids du mélange sont 0 ou 1 en fonction des individus, Lander et Botstein (1989) ont affirmé que la distribution asymptotique du processus de LRT, $\Lambda_{(.)}$, sur l'intervalle $[0, T]$, était le carré d'un processus d'Ornstein-Uhlenbeck. Ce résultat a été prouvé par Cierco (1998). Des bornes pour la distribution du maximum d'un processus d'Ornstein Uhlenbeck régularisé ont été proposées par Azaïs et Cierco-Ayrolles (2002), Azaïs et Wschebor (2009). Des résultats sur la distribution asymptotique du processus de LRT sous l'hypothèse nulle sont présentés dans Rebaï et al. (1994) pour une modélisation particulière des poids du mélange. Ces derniers résultats s'appuient sur les bornes données par Davies (1977), Davies (1987) pour le maximum de processus Gaussiens suffisament réguliers et pour des processus de chi-deux.

Dans notre étude, sont considérés les poids exacts utilisés par les généticiens pour la

détection de QTL. Tout d'abord, nous donnons la distribution asymptotique du processus de LRT sous l'hypothèse nulle d'absence de QTL sur $[0,T]$ ($q=0$) et sous l'alternative qu'il existe un QTL en $t^\star \in [0,T]$, ce qui signifie que pour chaque individu, le caractère quantitatif est distribué comme le mélange présenté en formule (1) en considérant $t = t^\star$. Par la suite, nous calculons la distribution asymptotique du processus de LRT sous l'hypothèse générale qu'il existe m QTL sur $[0,T]$ à $t_1^\star, \cdots, t_m^\star$ avec des effets additifs $q^1, \cdots, q^m$. La réponse est désormais un mélange de $M = 2^m$ composantes. Plus précisément, ce mélange est de la forme :

$$\sum_{\alpha=1}^{M} p_\alpha f_{(m_\alpha,\sigma)}(.)$$

où les p_α e les m_α sont des fonctions connues dépendant des paramètres inconnus μ, m, $t_1^\star$, ..., $t_m^\star$, q^1, ..., q^m. Sous cette alternative générale, le processus de LRT converge vers un processus Gaussien dont la fonction moyenne dépend de ces paramètres inconnus. Nous proposons d'estimer ces paramètres inconnus par vraisemblance pénalisée.
Enfin, nous démontrons que le processus de LRT est asymptotiquement le carré d'un processus d'interpolation non linéaire (i.e. la statistique de LRT à chaque point se déduit facilement des statistiques de Wald ou du score calculées aux positions où l'information auxiliaire est disponible).

Bibliographie

[1] Azaïs J. M., Cierco-Ayrolles C. (2002), An asymptotic test for quantitative gene detection, Ann. I. H. Poincaré.
[2] Azaïs J. M., Wschebor M. (2009), Level sets and extrema of random processes and fields, Wiley, New-York.
[3] Cierco C. (1998), Asymptotic distribution of the maximum likelihood ratio test for gene detection, Statistics.
[4] Davies R.B. (1977), Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika.
[5] Davies R.B. (1987), Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika.
[6] Rebaï A., Goffinet B., Mangin B. (1994), Approximate thresholds of interval mapping tests for QTL detection, Genetics.
[6] Wu R., Ma C.X., Casella G. (2007), Statistical genetics of quantitative traits, Springer.
[7] van der Vaart, A.W. (1998), Asymptotic Statistics, Cambridge series in statistical and probabilistic mathematics.